Towards a model for automatic action recognition for social robot companions

Towards a model for automatic action recognition for social robot companions

Ingo Keller*[1], Markus Schmuck[1], Katrin Solveig Lohan*[1]

***Abstract*— In this paper, we will explore human movement using a tutoring spotter system which controls an iCub robot. We will present an evaluation based on the captured human movement from an experimental study, where our participants demonstrated a salt-shaker and a cup-stacking task to the iCub robot. We will use a method of action recognition, which will help the robot to differentiate between these actions, as it will focus the robot's attention on the vital information presented by the human. Our findings imply that the behaviour of the robot affects our participants and it influences both, presentation time as well as the ratio between action and sub-action during the task presentation. Furthermore, the stability of the action recognition system is influenced by this modification of the human presentation.**

I. Introduction

Human-robot interaction is facing several barriers whilst moving out of the lab environment into the real world. One of these barriers is to detect, understand and learn from natural interaction with humans. In this paper, we will try using existing techniques on action recognition, in combination with natural human teaching, to see how far away we are from overcoming the barrier to learn from simple task presentations.

In the review of human activity analysis by Aggarwal and Ryoo [1], they classify activity detection approaches into two different classes. The first which are consisting of single-layered approaches, are based on sequences of images to recognise human activities. These are suitable for the recognition of gestures and actions with sequential characteristics. The second class are hierarchical approaches which are more suitable for high-level human activities that are composed of simpler human activities, e.g. multi agent interactions [6] or observing sport activities [15]. As we are interested in the learning from natural human tutoring which could include gestures as well as actions, we will select an application of the first class.

Using such a model for automatic action recognition for social robot companions, we explored the action presentations in a human-robot interactions study. Recent work in action recognition and classification presented by Wen et al. [4] proposes a system for an automated method. In this paper we are following their suggested approach and applying it to real experiments. This work is particularly interesting for us as it allows time independent clustering of action parts (action lets). Moreover, the data we are interested in is presenting objects and the tasks that can be carried out with this objects. Here we are looking at a table top scenario with small one-handed movements within a range of 20cm×40cm×30cm. Therefore, we are looking for an algorithm that allows us to go into detail. This seems to be provided by the one from Wen et al. [4].

The data used in this research is based on the idea of the advantage of a robotic learner when using developmental strategies for controlling the behaviour of the robotic system [2]. The iCub robot was used with two different controlling mechanisms for its feedback strategies [10]. The interaction between the human and the robot is structured by the human tutor showing the robot how to handle the motionese objects [13].

The object demonstrations are particularly interesting to learn and classify from a robot perspective, as they encode further information about object presentations. These objects can be classified as more manner or more path oriented objects [8]. This classification could help the robot later on to build up a concept of these action properties.

II. Study

A study to compare the differences in the tutoring behaviour of several human tutors towards the iCub robot within two different behaviour constraints was conducted. In one condition, the robot was controlled by a contingent reaction pattern and, in the other condition, the robot's behaviour was random.

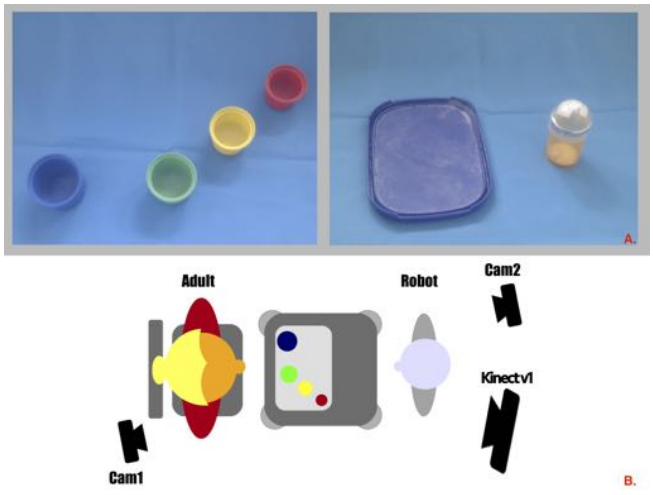

Fig. 1. Objects for task presentation and setup of our experimental study.

In the study, the participants (see Table I) were asked to present objects/tasks to the iCub robot. Due to missing data we had to exclude 4 participants 2 from each condition.

[1]Heriot-Watt University, MACS, Edinburgh

*Both authors Ingo Keller, as well as Katrin Solveig Lohan are corresponded authors as both have made equal afford on creating this work.

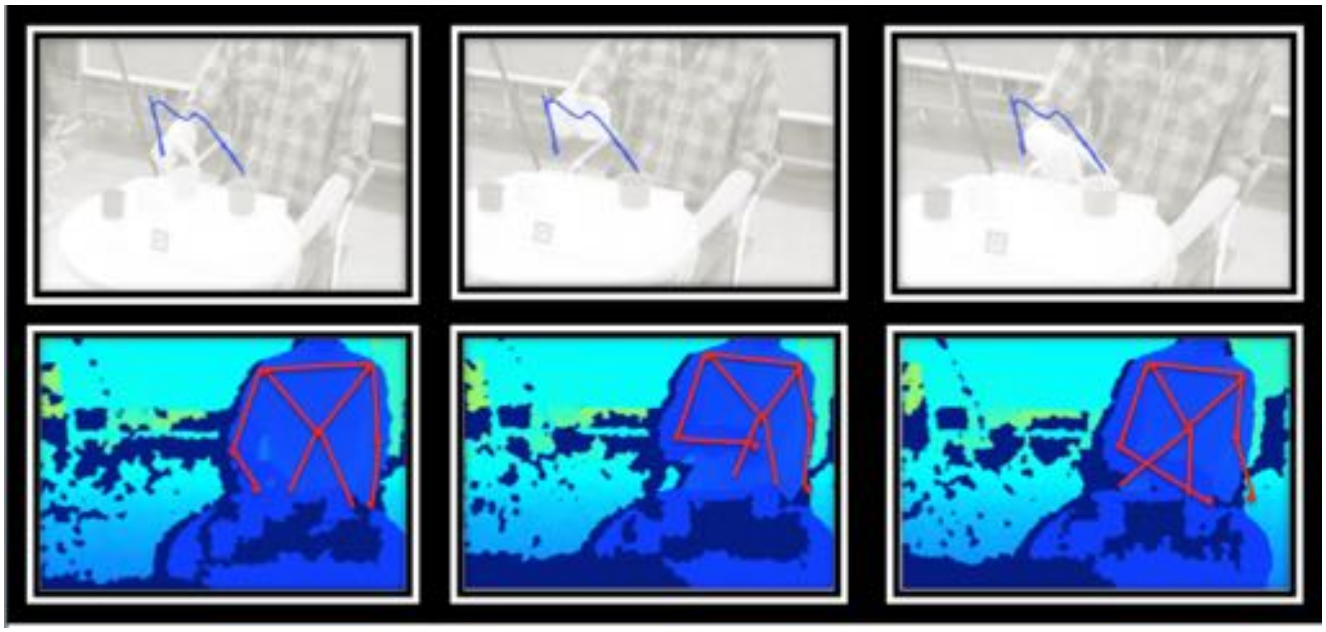

Fig. 2. The top pictures present the trajectory of one sub-action in the cup-stacking task it begins when the first cup is getting picked up by the human and ends when the cup is released into the next bigger one. The bottom pictures present the skeleton-tracking and 3D point cloud data from the kinect presenting the same sub-action.

Demographic — Behaviour	Contingency	Random
Total number participants	19	19
Participants age	19-68 years; median= 24 years	20-55 years; median = 25.5 years
Gender of participants	7 male; 12 female	9 male; 10 female
Participants with children	3	3
Estimated age of the robot	1 - 10 years; median= 5 years	0.3 -12 years; median = 4 years

TABLE I

SUMMARY OF BACKGROUND OF STUDY PARTICIPANTS.

Task — Behaviour	Contingency behaviour	Random behaviour
Manner oriented task	Contingency, Manner	Random, Manner
Path oriented task	Contingency, Path	Random, Path

TABLE II

2X2 DESIGN OF THE EXPERIMENT.

As defined by [13] we used objects were most of their functions and the objects themselves were familiar to the participants as they were either from a child's play chest or household objects like a salt-shaker (see Fig.1,A.). In total they were asked to present the use of 6 objects to the robot. In this paper we will present results based on 2 of the 6 objects, which were used in this experiment. In the beginning of the study participants had to present the same object/task, the lamp, to the robot. Due to the lack of randomess, the lamp object/task has been excluded from the evaluation. All other objects/tasks were randomised to avoid training effects on the participants. The objects/tasks presented by the participants are the following:

- Lamp task: Show the robot how to switch a table lamp on and off by pulling on a cord.
- **Cup-stacking task**: Stack cups into the blue cup in the following order: green, yellow, red.
- Minihausen task: Recreate a certain building block configuration by adding several blocks to a prearranged building block foundation.
- Bell task: Show the robot how to use a table bell.
- Ring task: Put three small plastic rings into a box.
- **Salt-shaker task**: Explain how a salt shaker could be used to pour some salt onto a blue lid.

In this paper we will concentrate our research on the **Salt-shaker** and the **Cup-stacking** task, the applied objects can be seen in figure 1,B.. This two objects/tasks were selected, as they both represent a unique group of tasks. More information about the selection of these two objects/tasks can be found in section V. Furthermore on these objects, previous research had been carried out using them. This includes motion analysis with different interaction partners, e.g. children, parents and other robots ([13], [10], [8], [12] and others), which gives us the opportunity to reuse manual annotations as well as get a rough understanding of the behaviour variation on these tasks. We selected 2 objects/tasks with repetition of the sub-action in the presentation of the human (see Fig. 2). Repetition helps our algorithm, as it makes the action segments similar and the data set gets multiplied by the amount of repetition for the input into the SVM. Overall there are 3 sub-actions for the cup-stacking task, such that we have $34 \times 3 = 102$ sub-actions as input from the SVM and we have up to 2 sub-actions in the salt-shaker task a total of 40 sub-actions.

The participants were tracked with a Kinect body tracking system. We collected 3D data representing skeleton-tracking from the human (see Fig. 2). The robot would behave either contingent or non-contingent to the participant. Therefore, we had a 2×2 design in the experiment, discussed below Table II.

The robotic system with the contingency condition was

reacting to the objects and the users gazing behaviour. In the non-contingency condition the robot exhibited random movement (see section IV).

III. Contingent Tutor Spotter

The tutoring spotter system was created as part of the ITALK project and is capable of providing socially contingent feedback using eye gaze and pointing [10], [9].

a) Gazing feedback: The robot detects three gazing classes (gazing at the robot, gazing away from the robot and gazing at an object) in the human tutor's behaviour and responds with the same behaviour. Hence, the robot would look at you if you are looking at it, the robot would look around the room if you are looking somewhere else and when you are looking at the object, the robot would follow your gaze to the object (see Fig. 3).

b) Pointing feedback: Child-directed action, and in particular object presentation, has been shown to facilitate learning in a tutoring situation. In particular, Matatyaho and Gogate [11] found that the demonstrating action, in which a tutor moves an object towards a student's face, is likely to produce a novel word-object relationship [5] and thus serves as a reliable method of learning words. Using the tutor spotter, the iCub robot responds to a demonstrating gesture by pointing towards the object that the tutor is moving (see Fig. 3).

IV. Object-driven Contingency

The robot's behaviour was based on tracking the objects or the face of the participant. As in the tutor spotter condition, the robot was able to look at the participant's face, at the object or somewhere else and it was able to use pointing gestures. The object-driven implementation, which focuses on tracking objects or, if no object is available, switches the robot's gaze to the tutor's face, seems to correspond to infants' behaviour [3].

a) Gazing feedback: The robot's gaze was controlled by a 'boredom' filter. If the same face or object was shown for too long, the robot switched to random gaze. This means that the gazing behaviour of the robot was based on timing of previous gazing behaviour.

b) Pointing feedback: The robot tracked the object and occasionally (on a random basis) pointed at the object. The robot did not use the tutor's social behaviour in this condition.

V. Manner vs Path oriented behaviour

In order for artificial intelligent systems to interact naturally with human users, they need to be able to learn from instructions when actions should be imitated. Human tutoring typically consists of demonstrations accompanied by speech. When demonstrating actions, humans show a distinction between two kinds of motion events: path-oriented actions and manner-oriented actions. These two kinds of actions are described in language by more path-oriented or more manner-oriented utterances. In path-oriented utterances, the source, trajectory or goal is emphasised whereas, in manner-oriented utterances, the medium, velocity or means of motion are highlighted. How this influences the demonstration for children on the means of gazing behaviour and language used has been researched by Lohan et. al. (2014) [8]. The findings related to the development of manner and path concepts have been used to implement new effective feedback strategies in the tutoring spotter system, which should help improve human-robot interaction.

We will explore the human movement towards this tutoring spotter system. We will present an evaluation based on the captured human movement presenting one manner (Salt-shaker task) and one path oriented object (Cup-stacking task) to the iCub robot. Our model of action recognition, which will be presented in the next section, shall help the robot to differentiate between manner and path oriented actions, which will focus it's attention on the vital information presented by the human.

VI. Action Classification

For action classification we followed the approach of Wen et al. [4] with a few adjustments as shown in this section. For comparison of different actions they propose a Spatio-Temporal Feature Chain (STFC) instead of using Dynamic Time Warping (DTW). They argue that DTW comes at the cost of a high probability for temporal misalignment which degrades classification performance. An STFC is supposed to be a duration independent representation of human actions which is based on 3D point trajectories of joint positions.

Instead of using video based position estimators for finding the joint positions, we established the positions using the Microsoft Kinect Sensor. By combining RGB and depth map information, the device was able to capture 3D joint positions in real time. From this data we were able to calculate the corresponding STFCs. Having a duration independent representation allows for using fixed input size classification methods such as SVM or kMeans.

A. Dataset and Preprocessing

Our dataset is composed of: the captured joint positions from the Kinect sensor and hand annotated action segmentation, which is based on the front view video sequences.

The Kinect sensor data provided information from the head, left hand, right hand and chest joints. Since our experiment setup involved the participant sitting in front of a table, which blocks parts of the chest, the chest sensor readings were not reliable and so were excluded from the following calculations. This problem somewhat occurs for the hand joints, as well resulting in the loss of position information in some frames. While this is unfortunate, it is not expected to interfere with the action classification. If the hand was hidden by the table, it means that it was not a part of the actions we used in our dataset and would be disregarded anyway, as we shall see later in the article.

For the action segmentation the videos were annotated using the ELAN software. In the Cup-stacking task an action is defined as: putting one cup into another one. The action

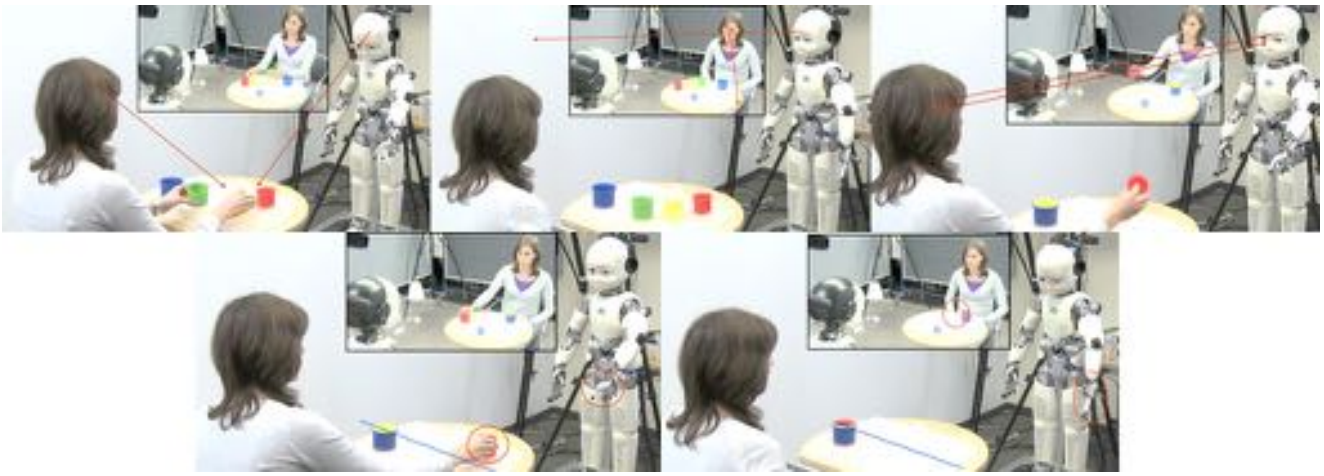

Fig. 3. These first three pictures show the gazing classes which were detected by the tutoring spotter. Top left: looking at the object, top right: looking at the interaction partner (iCub), top middle: looking somewhere else. The last two pictures illustrates the presenting and non-presenting class. Left bottom: presenting, right bottom: non-presenting class.

starts when the participant first touches the cup and ends when the participant places the cup into another one which resembles the notion of object transfer. Since there are three cups to be stacked, three actions are produced by each participant. For the Salt-shaker task the action starts when the salt shaker is picked up and stops when the shaker is again in an upright position. If the participant decided to pour the salt shaker again it would be regarded as another action. While only one action is expected more than one can occur.

Since the data was captured through different methods we had to synchronize it. We chose the timestamps from the Kinect sensor as a baseline. Each timestamp corresponds to one frame. Thus, we selected the corresponding frame for the hand annotated segmentation by choosing the one with minimal time difference.

B. Motion Features

Discriminating different actions from each other or aligning the same action from different people cannot be done using the positional information as it is. Therefore, we needed to determine the position and viewpoint independent information which can be used to compare movement patterns. This information is called *motion features* and can be calculated as shown below.

Our joint point trajectory is a parameterized matrix of joint point features per frame:

$$R = [p_1, p_2, \ldots, p_n] \text{ where } p_t = [\tau_t, x_t, y_t, z_t, v_t, \Delta h_t, \Delta\phi_t, \Delta\theta_t, k_t]^T$$

τ_t is the timestamp and x_t, y_t, z_t are the spatial coordinates at frame t where t is the index of the frame as captured by the Kinect sensor. These values are used for calculating the feature values, however in relation to the classification they are not regarded as motion features due to their positional information. For ease of notation we use $xyz_t = [x_t, y_t, z_t]$.

$v_t = \frac{xyz_t - xyz_{t-1}}{\tau_t - \tau_{t-1}}$ is the motion velocity at frame t and $v_0 = 0$ in the first frame. In contrast to [4] the velocity is calculated from the actual timestamps instead of using the frame index t. Due to the way the sensor captures the information the time between frames is not equidistant which needs to be taken into account.

$\Delta h_t = z_t - z_{t-1}$ is the change of height at frame t with $\Delta h_0 = 0$ in the first frame and shows movement in terms of moving up or down. $\Delta\phi_t$ and $\Delta\theta_t$ are the directional features at frame t, denoting left-right and further-closer information. While these features are supposed to be calculated from an egocentric perspective using the hip center as the center of a Spherical Coordinate System we needed to choose a different center point for which we used the Kinect sensor origin. This change of perspective is justified in this case by the fact that the participants did not change position as they sat in a chair throughout the experiment. Hence, a stable spatial relationship was maintained. The Spherical Coordinate System is mapped by (r, θ, ϕ), where θ is the inclination from the z-axis, ϕ is the azimuth from the x-axis in the xy-plane, and r is the radius [4].

$$k_t = \frac{\sqrt{(z_t'' y_t' - y_t'' z_t')^2 + (x_t'' z_t' - z_t'' x_t')^2 + (y_t'' x_t' - x_t'' y_t')^2}}{(x_t'^2 + y_t'^2 + z_t'^2)^{\frac{3}{2}}}$$

k_t is the curvature of a joint point at frame t as defined by equation VI-B. We determined the curvature by considering each of the x,y,z-trajectories as a smoothed curve in 3D space. These curves can be calculated using the Univariate-Spline implementation of the scipy software package which also provides the derivatives of the spline as needed to calculate the curvature.

Using the motion features, points of changing behaviour can be determined. These are called Segmentation Points (SP) and they denote the start or the end of an Actionlet (AL). Points between each two SPs are considered to have the same motion direction. A segmentation point is defined as $s = t$ iff $(\Delta h_t \to 0 \vee \Delta\phi_t \to 0 \vee \Delta\theta_t \to 0) \wedge k_{t-1} < k_t < k_{t+1}$ where $\to$ denotes for crossing 0.

$S = \{s_1, \ldots, s_m\}$ is the sequence of segmentation points where s_1 is the first segmentation point in the sequence and s_m the last. The value of s_i is the frame index in which the SP appears. $A = \{(s_1, s_2), \ldots, (s_{m-1}, s_m)\}$ where $s_i \in S$ are the Actionlets of the SP sequence.

A problem with motion analysis is that captured data from human motions usually contain some noise. On one hand, this noise might have been introduced by sensors due to their imperfect precision and jitter. On the other hand, every human motion comes with a small tremor. This tremor is further influenced by, for example, stress or health conditions.

To cope with this kind of noise the segmentation points get clustered. We utilized the *Hierarchical clustering* implementation of the scipy package. As cluster type we chose euclidean distance and created clusters with a distance $d = 10mm$. While that distance might seem high, lifting of objects can increase the tremor.

An AL-Graph $G = (V,E)$ gets constructed as follows:

$$V = \{v_i | v_i \subset S \wedge v_i \cap v_j = \emptyset \text{ for } i \neq j, 1 \leq i,j \leq n\}$$

$$E = \{e_j | e_j = < s_{j1}, s_{j2}, v_{j1}, v_{j2} >, s_{ji} \in v_{ji}, 1 \leq j \leq l, i \in \{1,2\}\}$$

v_i contains all SPs that are correlated to one cluster denoting a sphere in space with a maximum diameter of 10 mm. To maintain the time-ordering of the vertices the cluster labels are reordered so that the ordering of i corresponds to the ordering of $min(s_j) \in v_i$.

To remove the tremor and jitter we removed all edges in which the start and the end segmentation point is within the same cluster ($E = E \setminus \{e_i | v_{j1} = v_{j2}\}$).

C. Classification

The Spatio-Temporal Feature Chain is a sequence of motion feature nodes $\text{STFC} = [node_1, \ldots, node_i, \ldots node_{l_{setup}}]^T$. Each node is a feature vector of the selected 3D joint positions $node_i = [v_t, \Delta h_t, \Delta\phi_t, \Delta\theta_t, k_t], i = 1, \ldots, l_{setup}$. Features for the nodes are extracted by removing the positional information from the trajectories ($T : p_t \rightarrow [v_t, \Delta h_t, \Delta\phi_t, \Delta\theta_t, k_t]$). Only the joint with the highest velocity at one point gets used for further calculations. This is done by comparing $max(v_t)$ for all joints.

Wen et al. argue that there cannot be a template action pattern acquired from humans since "people's movements differ in thousands of ways ..." [4]. Therefore, we follow their strategy and compare actions by means of STFC with a fixed number of motion features for each action. They determined the length of the STFC ($l_{setup} = 30$) from observation of their datasets. Considering our dataset we appropriately used $l_{setup} = 30$ which provided the most overall recognition rate. The points are adjusted according to their method. For action classification we used the Support Vector Machine (SVM) implementation of the sklearn package which is based on libSVM. Following Wen et al. we chose the dot product as our kernel function for the SVM.

VII. Results

Figure 4 shows our evaluation results which includes the recognition rate of the system. All of our results were calculated using the following values: $l_{setup} = 30$, $d = 10mm$ (distance). The data has been split to show 50% for the trainings set and 50% for the test set. From this 51 data sets for the cup-stacking task and 20 data sets from the salt-shaker task. The value of the $l_{setup} = 30$ was selected based on the paper of Wen et al.. In our case this results in that we have to see 30 data points (which equals 30 frames) as the minimum length of an action. We selected a distance of 10 mm as we found this is an appropriate accuracy based on the noise of the sensors, the distance from the camera and the scale of the task at hand which is much larger than 10 mm. Furthermore, it was selected as the lifting of objects might increase the tremor.

Property Robot Behaviour	N	Mean	SD	SE	F	sig.
$\|R\|$ contingent	26	130.92	68.57	13.45	4.62	.04*
$\|R\|$ non-contingent	25	97.08	39.40	7.88		
$\|S\|$ contingent	26	27.04	15.14	2.97	3.95	.05*
$\|S\|$ non-contingent	25	19.52	11.56	2.31		
$\|AL\|$ contingent	26	26.04	15.14	2.97	3.95	.05*
$\|AL\|$ non-contingent	25	18.52	11.56	2.31		
$\|V\|$ contingent	26	9.15	5.46	1.07	1.18	.28
$\|V\|$ non-contingent	25	7.88	2.21	0.44		
$\|E\|$ contingent	26	12.08	7.40	1.45	8.07	.007**
$\|E\|$ non-contingent	25	7.68	2.30	0.46		

TABLE III

Results of the one-way ANOVA for the Cup-stacking task. N is based on the amount of STFC which could be calculated from our 102 data sets. Therefore only 50% of data sets had enough data points for the STFC to be calculated.

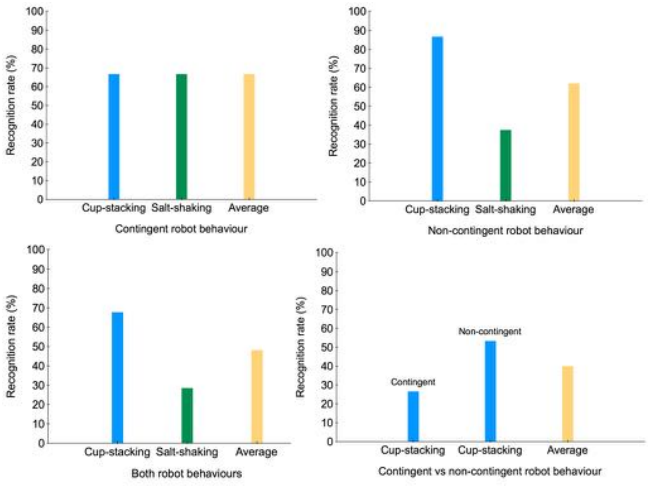

Fig. 4. Recognition rates. Top left: both tasks, when the participant was addressing the contingent robot. Top right: both tasks, when the participant was addressing the non-contingent robot. Bottom left: the combined data sets from all participants. Bottom right: if the system has seen the Cup-stacking task for both sets of participants (contingent vs non-contingent robot behaviour).

These results suggest that the robot's behaviour has an impact on the participants behaviour and therefore influences the recognition capability of our system. The top graphs in figure 4 show that when the robot is behaving in a contingent manner the recognition rate appears to be more consistent than in the case when the robot behaves in a non-contingent manner.

Looking at the bottom graphs, we can see that our system

Property Robot Behaviour	N	Mean	SD	SE	F	sig.
$\|R\|$ contingent	13	239	180.86	50.16	.59	.45
$\|R\|$ non-contingent	16	194.19	132.33	33.08		
$\|S\|$ contingent	13	40.77	30.91	8.57	1.40	.25
$\|S\|$ non-contingent	16	29.50	20.27	5.07		
$\|AL\|$ contingent	13	39.77	30.91	8.57	1.40	.25
$\|AL\|$ non-contingent	16	28.50	20.27	5.07		
$\|V\|$ contingent	13	17.23	11.60	3.22	.47	.50
$\|V\|$ non-contingent	16	14.44	10.32	2.58		
$\|E\|$ contingent	13	23.08	18.12	5.03	1.12	.30
$\|E\|$ non-contingent	16	17.06	12.38	3.10		

TABLE IV

RESULTS OF THE ONE-WAY ANOVA FOR THE SALT-SHAKER TASK. N IS BASED ON THE AMOUNT OF STFC WHICH COULD BE CALCULATED FROM OUR 40 DATA SETS. THEREFORE ONLY 72.5% OF DATA SETS HAD ENOUGH DATA POINTS FOR THE STFC TO BE CALCULATED.

appears to be sensitive to variation in the action presentation portrayed by the different robot behaviour (left graphic).
To explore these results in more depth, we went back to the properties of the point trajectories and their segmentations. We looked at: the number of points in R ($|R|$); the number of segmentations points in S ($|S|$); the number of actionlets in AL ($|AL|$); the number of clusters in V ($|V|$) and the number of edges in E ($|E|$). We compared the results of these properties for each condition (contingent vs non-contingent robot behaviour) by calculating a one-way ANOVA (see Table III and IV).

The results of this analysis suggest that there are significantly more data points, segments, edges and actionlets found in the Cup-stacking task, when the robot is demonstrating contingent behaviour towards the participant. However, there is no significant difference in the amount of clusters found. These findings imply that participants take longer to present the Cup-stacking task when the behaviour of the robot is contingent in it's behaviour. The participants create more sub-actions, e.g. looming motions when explaining the Cup-stacking task towards a robot that is behaving contingently to them. These findings agree with previous results on the tutor spotter system and imply that the motionese behaviour is induced using contingent robot behaviour (see [7], [14]).

VIII. FUTURE WORK

While we, found human motion patterns change according to the robots behaviour, the overall recognition rate for the actions was not that high as expected. However, this can be attributed to the relativly similar action pattern for both task in means of the AL-Graph. Hence, we would like to add more diverse motion patterns to our dataset. In the future we could include more than one joint for classification. In order to be able to include all joints we have to find a way to deal with missing data. This is not only useful for our own dataset in cases of missing frames but also for all real world tasks in which occlusion can occur. We would like to validate our method using benchmark datasets, especially since we used a more restricted set of motion features. An issue that has not been addressed so far is the segmentation of continuous data streams of 3D point trajectories. Since the notion of an action is not yet well-defined, we need to find an automatic measure to determine the start and the end of an action.

ACKNOWLEDGMENT

The authors would like to thank Isabel Wessemann for her help with executing the user study. We would also like to thank Christian Dondrup for his assistance with the robotic platform and all the participants that kindly agreed to interact with our iCub. We also appreciate the financial support provided by the ITALK-Project and by Heriot-Watt University. Finally, we are grateful for the guidance provided by Katharina Rohlfing and Britta Wrede.

REFERENCES

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Angelo Cangelosi, Matthew Schlesinger, and Linda B Smith. *Developmental robotics: From babies to robots*. MIT Press, 2015.

[3] Kaya de Barbaro, Christine M. Johnson, Deborah Forster, and Gedeon O. Deak. Temporal dynamics of multimodal multiparty interactions: A microgenesis of early social interaction. In A.J. et al. Spink, editor, *Proceedings of Measuring Behavior 2010*, pages 247–249, Eindhoven, The Netherlands, 2010.

[4] Wenwen Ding, Kai Liu, Fei Cheng, and Jin Zhang. STFC: Spatio-temporal feature chain for skeleton-based human action recognition. *Journal of Visual Communication and Image Representation*, 26:329–337, January 2015. 00002.

[5] L.J. Gogate, L.H. Bolzani, and E.A. Betancourt. Attention to maternal multimodal naming by 6-to 8-month-old infants and learning of word–object relations. *Infancy*, 9(3):259–288, 2006.

[6] Stephen S Intille and Aaron F Bobick. A framework for recognizing multi-agent action from visual evidence. *AAAI/IAAI*, 99:518–525, 1999.

[7] Katrin S Lohan, Katharina Rohlfing, John Saunders, Chrystopher Nehaniv, and Britta Wrede. Contingency scaffolds language learning. In *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[8] Katrin Solveig Lohan, Sascha Griffiths, Alessandra Sciutti, Tim Partmann, and Katharina Rohlfing. Co-development of manner and path concepts in language, action and eye-gaze behavior. *Topics in Cognitive Science*, 2014.

[9] Katrin Solveig Lohan, Karola Pitsch, Katharina Rohlfing, Kerstin Fischer, Joe Saunders, Hagen Lehmann, Chrystopher Nehaniv, and Britta Wrede. Contingency allows the robot to spot the tutor and to learn from interaction. In *ICDL-EPIROB 2011*, 2011.

[10] Katrin Solveig Lohan, Katharina Rohlfing, Karola Pitsch, Joe Saunders, Hagen Lehmann, Chrystopher Nehaniv, Kerstin Fischer, and Britta Wrede. Tutor spotter: Proposing a feature set and evaluating it in a robotic system. *International Journal of Social Robotics*, 4:131–146, 2012.

[11] D.J. Matatyaho and L.J. Gogate. Type of maternal object motion during synchronous naming predicts preverbal infants' learning of word–object relations. *Infancy*, 13(2):172–184, 2008.

[12] Yukie Nagai and Katharina J Rohlfing. Computational analysis of motionese: What can infants learn from parental actions. In *Proc. Int. Conf. on Infant Studies (ICIS 2008)(March 2008)*, 2008.

[13] K.J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006.

[14] Anna-Lisa Vollmer, Katrin Solveig Lohan, Kerstin Fischer, Yukie Nagai, Karola Pitsch, Jannik Fritsch, Katharina J Rohlfing, and Britta Wredek. People modify their tutoring behavior in robot-directed interaction for action learning. In *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, pages 1–6. IEEE, 2009.

[15] Elden Yu and Jake K Aggarwal. Detection of fence climbing from monocular video. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 375–378. IEEE, 2006.